Weekly Report
12/01/2014 - 12/07/2014

Jing Xia

December 8, 2014

1 Summary

This week I mainly focused on the rank project, the data inspection project and the driving class.

2 Projects

2.1 Project 1 - Rank Visualization

I've fixed the dataset problem and now the 2014 dataset is complete from January to September. Also, I've implemented fetching semantic information from DBpedia. But the query cannot get through unless I use a foreign proxy.
The appendix is the current draft of the paper, just abstract and introduction.

2.2 Project 2 - Data Inspection

1. Qualitative organization of collections of shapes via quartet analysis [1].

   This paper organizes a heterogeneous collection of shapes with a hierarchical categorization tree. The tree is constructed with a quartet forming, a three-dimensional sphere embedding and a MaxCut-like partitioning.

   We can apply a similar process to our view exploration. Initially our system provides an overview of a similarity graph of all possible views (one dimension or two dimensions). Users can further explore the dataset by selecting a subset. By such filtering, the view graph expands to a second layer (an imaging of the first layer but with filtered data), where users can zoom in to data of interest. Users can also swap between imaging views between two layers to compare the difference.

2. SEEDB: Automatically Generating Query Visualizations [3].

   SeeDB provides a solution to compare between two subset of data. And our data inspection project can learn from it.

Applying the methods described in the two papers above, I designed our data visual inspection (DataVIP) system as follows:

- **Layers of view graph** are explored progressively, so is the underlying data, by users selecting a sub-dataset from one view.
- For each layer, the view graph has **a graph layout** that keep similar views close and dissimilar views apart.
- For each view, the view ?node? is represented with **a glyph-based view descriptor** that characterizes the data distribution and its chart type.

2.3 Project 3 - NBA Game Visualization

-

3 Paper Reading

Please refer to Project 2.

4 Miscellaneous

This week I had three driving classes, each for a whole morning. Also, we had a chance to communicate with data analysts in Alibaba.

5 To Do List

1. TopK: adjustment of similarity measurement
2. TopK: A more detailed draft of paper.
3. NBA project discussion and data inspection project discussion.

References

[1] Shi-Sheng Huang, Ariel Shamir, Chao-Hui Shen, Hao Zhang, Alla Sheffer, Shi-Min Hu, and Daniel Cohen-Or. Qualitative organization of collections of shapes via quartet analysis. *ACM Transactions on Graphics*, 32(4):1, July 2013.

[2] Jock Mackinlay, Pat Hanrahan, and Chris Stolte. Show me: Automatic presentation for visual analysis. *Visualization and Computer Graphics, IEEE Transactions on*, 13(6):1137–1144, 2007.

[3] M Vartak, S Madden, and A Parameswaran. SEEDB: Automatically Generating Query Visualizations. In *Proceedings of the VLDB*, pages –, January 2014.

Appendices

A Abstract

Wikipedia top page view statistics are collections of top viewed Wikipedia pages over time, and of great importance in analyzing users' interest or current affairs. However, visualizing Wikipedia top page-view statistics usually suffers from great visual clutter. We formed a principle that any good design should connect the same Wikipedia page over time without causing unnecessary perceptual complexity. Following gestalt's law of continuity, we tried out a variety of visual designs of top ranked pages. Also, users are able to explore connections among those top viewed pages by taking both the page-view behavior and the page-link information into consideration. Such combination enhances the unweighted Wikipedia page-link network and brings users' page of interest in focus. We conducted a user study of the various visual designs and evaluated the usage of the system. The results show the feasibility of the visual design and the system.

B Introduction

Wikipedia is considered as the biggest online encyclopedia, everyday page-view can be more than 600M in all languages. It has become a knowledge exchange market where users learn and contribute their knowledge. Due to the accumulation, Wikipedia has also become a huge knowledge warehouse.

As opposed to various researches regarding infrastructures of Wikipedia such as NLP and data warehouse, work on time variance of Wikipedia top page-views is still new. According to the 20/80 principle, the ranking data of Wikipedia top page-view statistics (Wikipedia page rank for short) reflects users' major interests in Wikipedia or furthermore in current affairs. Time series of Wikipedia page-rank, or Wikipedia page-ranking trends therefore conveys how users' social interest evolving over time. However traditional data mining strategies does not satisfy users' various needs towards the long Wikipedia ranking series. On the other hand, visualization of Wikipedia page-ranking changes, which mostly involves continuous visual metaphors such as band or river, suffer greatly from visual clutter.

Our WikiTrend system not only visualizes Wikipedia page-ranking trends, but also constructs a semantic network based on a given Wikipedia page of interest. Our visualization design dodges visual clutter effectively by breaking band-like or river-like visualization into scattered glyphs while keeping users' perceptual continuity towards certain ranking items. On the other hand, the semantic network associates the underlying Wikipedia pages into a current affair. Moreover, it also characterizes a user-aware network of the original planar Wikipedia page-link network. Although only practiced on the Wikipedia dataset, such design can be applied to other time series such as stock prices. In summary, we summarize our contributions as follows:

- A novel glyph design that portrays the ranking trend of Wikipedia top viewed pages.
- A measurement that mines Wikipedia pages with similar ranking trends.
- A mashup of page-view and page-link that endows semantic information to Wikipedia pages similarity network.